LoRA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS

Introduction

- Fine-tuning pre-trained LLMs usually implies adapting all the weights of scale xx billions
- LoRA instead optimizes this by indirectly adapting low rank decomposition matrices while keeping the original weights frozen
- Number of trainable parameters can be reduced to 0.01%

Existing Solutions : Drawbacks

- Adapter Layers: Introduce inference latency due to sequential processing
- Prefix Tuning: Difficult to optimize and also decreases available sequence length

LoRA

- Limits to adaptation of only Attention weights in transformers
- Reduce hardware usage by 10000x when adapting only Wv matrix and rank=4 which facilitates switching of models quickly when deployed
- Increased training speed of 25%
- Performs significantly better than other alternatives like FineTuning(FT), Bias-only or BitFit, Prefix-embedding tuning, Adapter tuning

Understanding Low Rank Updates

- Adapting both Wq and Wv, each with r=4, yields the best results over just Wq with r=8
- Adaptation matrix A can have very low rank (potentially 1) given that top singular-vector directions is the most important
- New weight adaptation matrix W’ has strong correlation with original W as it amplifies directions that are not emphasized in W